SERIEs (2012) 3:291–310
DOI 10.1007/s13209-011-0074-3

ORIGINAL ARTICLE

The evaluation of citation distributions

Javier Ruiz-Castillo

Received: 2 December 2010 / Accepted: 30 June 2011 / Published online: 10 August 2011

Abstract This paper reviews a number of recent contributions that demonstrate that a blend of welfare economics and statistical analysis is useful in the evaluation of the citations received by scientific papers in the periodical literature. The paper begins by clarifying the role of citation analysis in the evaluation of research. Next, a summary of results about the citation distributions' basic features at different aggregation levels is offered. These results indicate that citation distributions share the same broad shape, are highly skewed, and are often crowned by a power law. In light of this evidence, a novel methodology for the evaluation of research units is illustrated by comparing the high- and low-citation impact achieved by the US, the European Union, and the rest of the world in 22 scientific fields. However, contrary to recent claims, it is shown that mean normalization at the sub-field level does not lead to a universal distribution. Nevertheless, among other topics subject to ongoing research, it appears that this lack of universality does not preclude sensible normalization procedures to compare the citation impact of articles in different scientific fields.

Keywords Citation analysis · Power law · Research performance · Poverty measurement · European paradox

JEL Classification O31 · Y80 · Z00

The author acknowledges financial support from Santander Universities Global Division of *Banco Santander*, as well as from the Spanish MEC through grant SEJ2007-67436. This paper is part of the SCIFI-GLOW Collaborative Project supported by the European Commission's Seventh Research Framework Programme, Contract no. SSH7-CT-2008-217436. This paper is the result of the author's joint work with Pedro Albarrán, Juan A. Crespo, Neus Herranz, and Ignacio Ortuño, whose comments and suggestions are gratefully acknowledged. Comments by Jaime Luque, Jaume Sempere, and Matthew Jackson are also greatly appreciated.

J. Ruiz-Castillo (✉)
Departamento de Economía, Universidad Carlos III and Research Associate of the CEPR Project SCIFI-GLOW, Madrid, Spain
e-mail: jrc@eco.uc3m.es

1 Introduction

There are different ways in which economists have approached the study of scientific activity. This paper focuses on some aspects of scientific performance that are readily observable, namely, citation distributions whose elements are the number of citations received by research papers published in the periodical literature. Scientists, including economists, may justifiably have reservations, even serious doubts, about the role of citation analysis in the evaluation of research. After summarizing what can be learned in this respect from the bibliometrics literature, the aim of the paper is to demonstrate that good data, sound statistical procedures and a certain dose of applied welfare economics are useful in pushing forward the state of the art in the evaluation of citation distributions. This is accomplished by reviewing some recent papers in this area (see Albarrán et al. 2010, 2011a,b,c,d; Albarrán and Ruiz-Castillo 2011), as well as some ideas and provisional results from ongoing research by a team that includes Pedro Albarrán, Juan A. Crespo, Neus Herranz, and Ignacio Ortuño.

It is well known that, due to vastly different publication and citation practices, the distributions of references made and citations received by scientific articles have very different characteristics across fields. For example, in the dataset that is described below, Economics and Business and Molecular Biology and Genetics have 62,685 and 150,237 articles, which represent 1.3% and 3.1% of the total number in all sciences published in 1998–2002. After a five-year citation window, the mean citation rate (MCR hereafter) and the h-index in Economics and Business is 3.3 and 64, almost six, and 4.3 times smaller than in Molecular Biology and Genetics where these statistics are 18.2 and 266.[1] In this context, it is not surprising that the evaluation of research units working in closely related but nevertheless heterogeneous sub-fields is usually carried on after a normalization procedure that takes into account differences in MCRs across sub-fields. The starting point of this paper is the observation that this diversity is compatible with the belief among scientometrics practitioners that citation distributions share some fundamental characteristics. In particular, it is generally believed that citation distributions are highly skewed, and it is widely held that citation distributions can be represented by power laws or Pareto distributions.[2] On the other hand, in an important recent contribution Radicchi et al. (2008) claim that since citation distributions only differ by a scale factor, after appropriate normalization we can speak of a universal citation distribution (see also Glänzel 2010). This would provide a solid grounding for the comparison of citations received by articles in different scientific fields.

As we see it, the problem is twofold. Firstly, the empirical evidence sustaining these empirical regularities is, although valuable, not conclusive (see Albarrán et al.

[1] The h-index, originally suggested by Hirsch (2005) to assess the citation impact of individual researchers, can be equally used to assess the citation performance of other research units and scientific fields. For example, the fact that the h-index in Economics and Business is 64 means that after a five-year citation window there are 64 articles published in 1998–2002 receiving 64 or more citations.

[2] An extensive discussion of the properties of power laws that have appeared in a variety of settings can be found in the reviews by Mitzenmacher (2004) and Newman (2005) and references therein. Egghe (2005) is a treatise on the importance of power laws for information production processes of which citation distributions are only one type.

2011d, for a review of the bibliometrics literature, as well as notes 2 and 6 in Jackson and Rogers 2007, referring to the lack of systematic and careful statistical testing of the key empirical regularities shared by socially generated networks). Secondly, the evaluation of research units' citation impact does not fully exploit the characteristics of citation distributions. This paper summarizes our contribution to a solution of these two shortcomings.

In the first place, using a large dataset of articles published in more than 8,000 academic or professional journals indexed by Thomson Scientific (TS hereafter), previously known as the Institute for Scientific Information (ISI hereafter), the following two facts are well established (Albarrán and Ruiz-Castillo 2011, and Albarrán et al. 2011d).

(1) Using size- and scale-independent descriptive statistics it is found that the shapes of reference and citation distributions are strikingly similar across a wide array of 219 sub-fields, identified with the Web of Science (WoS hereafter) categories distinguished by TS: references made by articles in any sub-field give rise to a highly skewed distribution of citations received, in which a large proportion of articles gets none or few citations while a small percentage of them account for a disproportionate amount of all citations.

(2) Using state-of-the-art maximum likelihood methods, we find that in 140 out of 219 sub-fields it cannot be rejected that a power law represents the upper tail of citation distributions. When they exist, power laws generally represent a small proportion of the upper tail of citation distributions but account for a considerable percentage of all citations. However, power laws' characteristics are subject to a large dispersion. Together with other evidence, this implies that the universality claim found in Radicchi et al. (2008) breaks down at both ends of citation distributions.

In the second place, we find that the skewness of citation distributions has important consequences for the evaluation of research units' performance. To begin with, a single statistic of centrality, such as the MCR or the median, may not adequately summarize these distributions for which the upper and the lower part are typically very different. In a first alternative, we suggest investigating the units' publication shares at every percentile of the world citation distribution in each field (Albarrán et al. 2010). However, the mere percentage of articles satisfying some interesting condition only captures what can be referred to as the *incidence* aspect of the phenomenon in question. A second alternative begins with the observation that, due to their skewness, it seems useful to describe a citation distribution by means of two real valued functions defined over the subsets of articles with citations above or below a *critical citation line* (CCL hereafter). These are referred to as a *high-* and a *low-impact indicator*, respectively (Albarrán et al. 2011a). Economists will surely recognize that the key to this approach is the identification of a citation distribution with an income distribution. Once this step is taken, the measurement of low-impact, which starts with the definition of low-citation papers as those with citations below the CCL, coincides with the measurement of economic poverty that, as originally suggested in Sen (1976) seminal contribution, starts with the definition of the poor as those individuals whose incomes are below a certain poverty line. In turn, once low-impact has been identified with economic poverty, it is equally natural to identify the measurement of high-impact with the measurement of a certain notion of economic affluence. In the first empirical

application of this methodology, Albarrán et al. (2011b) use a family of scale and replication invariant indices—originally suggested by Foster et al. (1984)—that satisfies a number of desirable properties, and has been widely used for the measurement of economic poverty in the last 25 years. These same properties lead to the selection of an equally convenient class of high-impact measures. Certain members of these two families of indicators are capable of simultaneously taking into account not only the incidence, but also what we call the *intensity*, and the *citation inequality* that affect the high-and low-impact phenomena they attempt to measure.

It should be noted that there is a number of indicators of citation excellence that fail to be scale- or replication-invariant but possess other interesting properties. This is the case of the h-index, an indicator that is robust to the presence of extreme observations in the form of articles with an extremely large number of citations—a property not satisfied by our high-impact indicator (for alternative characterizations of the h-index, see Woeginger 2008a,b; Marchant 2009; Quesada 2009, 2010, and for a survey of research on this index and its many variants see Alonso et al. 2009). The comparison of research units in terms of an index that is not scale and replication invariant poses formidable problems that are the subject of our current research.

The rest of the paper is organized in four sections. Section 2 briefly discusses the role of citation analysis in the evaluation of research. Section 3 introduces the notion of a homogeneous field, discusses the difficulties in finding its empirical counterpart in our database, and summarizes what is known about the typical shape of citation distributions, as well as the possibility of representing them by a power law at different aggregation levels. Section 4 is devoted to the evaluation of research units in terms of citation impact. In particular, it illustrates the use of a pair of high- and low-citation impact indicators in an important empirical problem: the comparison of the citation impact achieved in three geographical areas (i) the US, (ii) the EU, namely, the 15 countries forming the European Union before the 2004 accession, and (iii) all other countries in the rest of the world (RW hereafter). This is done in a convenient dataset where each article is assigned to only one of 22 broad fields distinguished by TS. Section 5 offers some concluding comments and suggests some possible extensions, part of which are the subjects of ongoing research.

2 What are citation counts good for?

Apart from a few pioneers and just after the publication in 1963 of the *Science Citation Index* by the ISI under the leadership of Eugene Garfield, the first systematic use of citations as a measure of impact, quality, and intellectual influence came out of Robert K. Merton's seminar at Columbia University during the late 1960s (see the references in Cole 2000). According to Merton's normative citation theory, citations represent intellectual or cognitive influence on scientific work. At the same time, a large literature has developed which holds that the probability of being cited depends on many factors that do not have to do with the accepted conventions of scholarly publishing, to say nothing of constructivist sociologists of science for whom the cognitive content of articles has little influence on how they are received. This is certainly not the place for an evaluation of these contending positions, very ably surveyed in

Bornmann and Daniel (2008), and discussed in van Raan (2004, 2005) and Weingart (2005). Instead, I would summarize my own position in the following three points (for a more detailed discussion, see the Working Paper version of this paper: Ruiz-Castillo 2011).

(1) The notion of scientific "quality" is virtually impossible to operationalize. The evaluation of the cognitive, methodological, and esthetic quality components of any research contribution can only be based on intrinsic scientific criteria assessed by qualified colleague researchers under the peer review system. However, communication is a crucial aspect of scientific endeavor, and members of the "invisible college" that is permanently discussing research results often play their role as critics by referring in their own work to earlier work of other scientists. Even if we remain agnostic about the myriad of citation motives researchers have, for our purposes it suffices to admit that, in principle, citation impact and citation distributions are worth investigating.

(2) Even the most fervent advocates of citation analysis would recognize that the citation process is a complex one that does not provide an ideal monitor of scientific performance. This is particularly the case at a statistically low aggregation level, i.e. the individual researcher or small institutions for which citation distributions tend to be small and, therefore, noisy from a statistical inference point of view. Consequently, in the sequel we will only refer to evaluation methods for entire scientific fields, or research units of a certain size, namely, a university department, research institute, journal, region, country, or supra-national geographic area.

(3) Bibliometric studies using citation counts are complementary to peer review judgments in at least two ways. (i) They may reveal macro-patterns in the communication process that cannot be seen from the limited perspective of the individual researcher. (ii) They may work as a control of peer review. When the results of the two evaluation exercises disagree, those responsible for peer review must provide an explanation, whereas when supported by bibliographic methods peer review judgments gain outside credibility. The conjunction of the two modes forms what Weingart (2005) calls "*informed peer review*", a commendable evaluation procedure to which we would like to be able to contribute.

3 The skewness and universality of science

3.1 Implementation problems

To examine whether citation distributions are similar or not, we must first confront what we should understand by a scientific field, and how it should be identified in practice. From an operational point of view, a scientific field is a collection of papers published in a set of closely related professional journals. A field is said to be *homogeneous* if the number of citations received by its papers is comparable independently of the journal in which each has been published. Consequently, if one paper has twice the number of citations as another in the same homogeneous field, it can be said not only that it has twice the international impact but also that it has twice as much merit as the other.

Naturally, the smaller the set of closely linked journals used to define a given research field, the greater the homogeneity of citation patterns among the articles

included must be. Therefore, ideally one should always work at the lowest aggregation level that the data allows. In the sequel, research areas at that level are referred to as *sub-fields*. In our case, this may mean the 219 WoS categories distinguished by TS. However, articles are assigned to WoS categories through the assignment of the journals where they have been published. Many journals are unambiguously assigned to one specific category, but many other typically receive a multiple assignment. As a result, only about 58% of the total number of articles published in 1998–2007 is assigned to a single WoS category. In this section, we deal with the problem of multiple assignments by means of a multiplicative strategy where each article is classified into as many sub-fields as WoS categories in the original dataset. An article assigned to three WoS categories, for instance, is classified into the three corresponding sub-fields and, therefore, it is counted three times. In this way, the space of articles is expanded as much as necessary beyond the initial size. As a matter of fact, the total number of articles in what we call the *extended count* for the 219 TS sub-fields is 57% larger than the original dataset.

Given the plethora of scientific sub-fields, for many practical problems the interest of investigating larger aggregates is undeniable. Above sub-fields, we distinguish between an intermediate category—referred to as *disciplines*, such as Internal Medicine or Dentistry; Particle and Nuclear Physics or Physics of Solids; and Organic or Inorganic Chemistry—and traditional, broad fields of study such as Clinical Medicine, Physics, and Chemistry, referred to simply as *fields*. For our purpose, it would be very convenient to have a hierarchical Map of Science organizing sub-fields, disciplines, and fields in a way agreed upon by the international scientific community. However, extreme doses of scientific inter-disciplinarity have it made impossible to count on such a Map (see Albarrán et al. 2011d, for some of the main references in this particularly active research field in Scientometrics). Given the difficulties inherent in any aggregation scheme, to climb up from the sub-field to the discipline and the field levels we use three alternatives routes.[3]

3.2 Empirical results

The question we investigate in this Sub-section is whether citation distributions are similar or not at the sub-field level, and whether the common features that are found are preserved in aggregation. As indicated in the Introduction, the evidence in the bibliometrics literature is very scant. Consequently, we have tried to set the record straight by investigating these issues with a large dataset consisting of about 3.7 million articles published in 1998–2002, the 97 million references they make, and the 28 million citations they receive after a common five-year citation window for every year, namely, from 1998 to 2002 for articles published in 1998, up to 2002–2006 for articles published in 2002. The 219 sub-fields include 77 in the Life Sciences, 36 in

[3] The first one, inspired by Tijssen and van Leeuwen (2003), distinguishes between 38 disciplines and 12 fields. The second one, inspired by Glänzel and Schubert (2003), distinguishes between 61 disciplines and 12 fields. The third one, constructed so as to maximize the appearance of a power law at the upper aggregation levels, consists of 80 disciplines and 19 fields.

the Physical Sciences, 73 in Other Natural Sciences, and 33 in the Social Sciences.[4] The main results in Albarrán et al. (2011d) can be summarized as follows.

3.2.1 Characteristics of reference and citation distributions

(1) Publication practices are very different indeed. In some research areas authors publishing one article per year would be among the most productive, while in other instances authors—either working alone or as members of a research team—are expected to publish several papers per year. This and other factors lead to reference and citation distributions which are very different in size: at the 219 sub-fields level, the mean is equal to 26,984 articles and the standard deviation is 29,669.

(2) Due to vastly different citation practices, reference distributions are very different across sub-fields. On average, the mean reference rate is equal to 26.3 and the standard deviation is 8. In turn, the ratio of references made over citations received is equal to 6.1 with a standard deviation of 4.1.[5] This is an important factor in explaining the dramatic changes experienced by the percentage of uncited articles and the MCR when we turn from reference to citation distributions: the first variable increases (up to 24.7%), while the second decreases (down to 5.7 citations) by a factor of five. Again, large standard deviations (13.9 and 3.5, respectively) indicate that citation distributions are very different indeed.

3.2.2 Characteristics of the shape of reference and citation distributions

(3) Size- and scale-independent descriptive tools permit us to focus on the *shape* of distributions. In particular, the Characteristic Scores and Scales (CSS hereafter) approach, pioneered by Schubert et al. (1987) in citation analysis, permits the partition of any distribution of articles into five classes according to the citations they receive. Denote by s_1 the MCR; by s_2 the mean of articles above s_1, and by s_3 the mean of articles above s_2. The first category includes articles without citations. As for the remaining four, articles are said to be *poorly cited* if their citations are below s_1; *fairly cited* if they are between s_1 and s_2; *remarkably cited* if they are between s_2 and s_3, and *outstandingly cited* if they are above s_3.

For the partition of reference and citation distributions at the sub-field level into three broad classes—comprising categories $1+2$, 3, and $4+5$—it is found that both the shape of reference distributions and the shape of citation distributions are strikingly similar. For brevity, it suffices to say that reference distributions are moderately skewed. However, as expected, citation distributions are highly skewed: approximately 69% of all articles receive citations below the mean and account for, at most, 21% of all citations, while articles with a remarkable or outstanding number of citations

[4] The six categories in Economics and Business are not very satisfactory: Agricultural Economics and Policy; Industrial Relations and Labor; Economics; Business; Business, Financial; and Management.

[5] Recall that references are made to many different items: articles in TS-indexed journals, as well as articles in conference volumes, books, and other documents, none of them covered by TS. Moreover, some references are to articles published in TS journals before 1998 and, hence, outside our dataset.

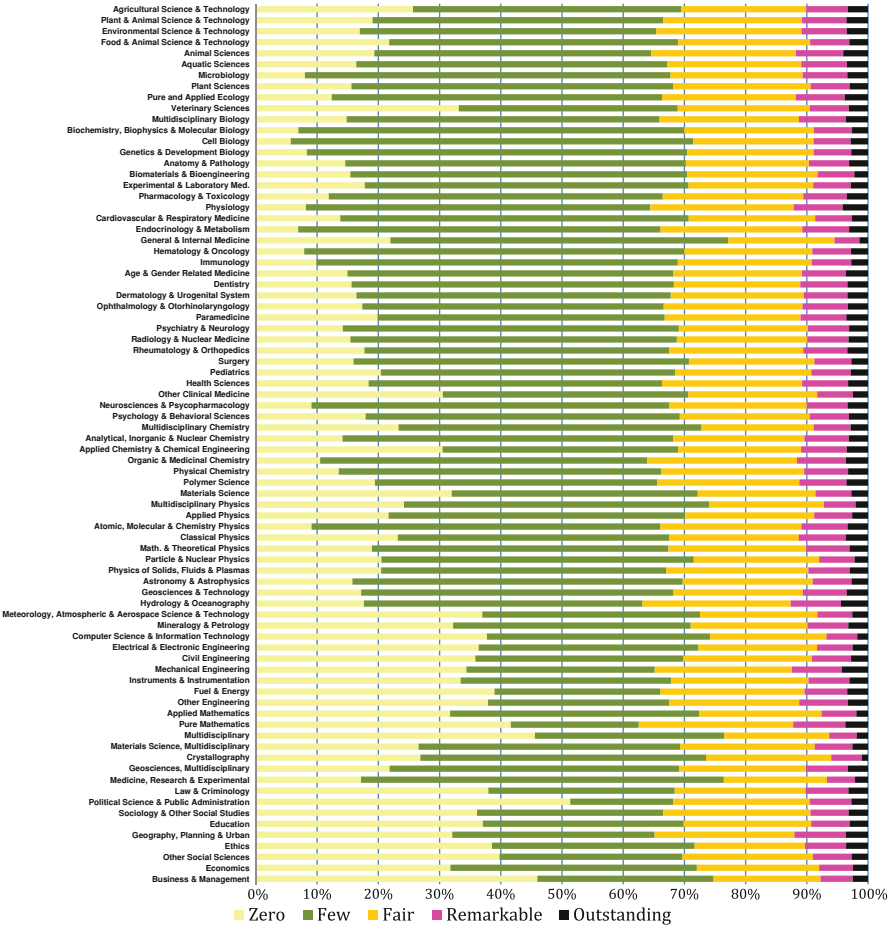

Fig. 1 Citations received by articles published in 1998–2002 with a five-year citation window

represent about 9% or 10% of the total, and account for approximately 44% of all citations.

Since sub-field shapes are so similar, any reasonable aggregation scheme should preserve its main characteristics. This is exactly what is found when sub-fields are aggregated into what we call disciplines and fields according to the three schemes mentioned in note 3 (see the dispersion statistics at all aggregation levels in Table 6 in Albarrán et al. 2011d). Figure 1, prepared for this paper, illustrates the classification of citation distributions into the five CSS categories for the 80 disciplines introduced in Section V.4 in Albarrán et al. (2011d).

(4) The partition into categories $1 + 2$, 3, and $4 + 5$ is, approximately, 70/20/10. However, when we move inside the union of categories 1 and 2 and categories 4 and 5 differences across disciplines become very large. Dispersion statistics formally reveal that the universality of citation distributions breaks down at both the lower and the upper tails at all aggregation levels [Waltman et al. (2011) reach the same conclusion

with a different methodology]. This conclusion contrasts with the more optimistic view offered by Radicchi et al. (2008) with a methodology that omits articles without citations, examines distributions at a limited set of points and, above all, covers only 14 of the 219 sub-fields.

3.2.3 The prevalence of power laws

(5) On the other hand, using maximum likelihood estimation methods (Clauset et al. 2007) it can be concluded that the existence of a power law representing citation distributions is a prevalent but not a universal phenomenon: in 140 out of 219 sub-fields, covering about 62% of the total number of articles in the sample, the existence of a power law cannot be rejected. However, when they exist, power laws (i) have a scaling parameter larger than usually believed (the median value is 3.85), with the implication that the citation inequality among the articles in the power law is smaller than what was previously believed, (ii) only represent a small proportion of the upper tail of citation distributions, and (iii) account for a considerable percentage of all citations. Although subject to a large dispersion, on average power laws represent 2% of all articles in a sub-field, and account for about 13.5% of all citations.

(6) When moving up from the sub-field level to other aggregate categories, we find that the power law algebra operates in a very subtle way: sub-fields for which a power law does not exist may be aggregated into a category for which the existence of a power law cannot be rejected. On the other hand, power law behavior at the sub-field level is not always preserved in aggregation; in particular, a single sub-field may be responsible for the power law behavior of a large number of sub-fields to disappear. Heterogeneous broad fields, such as Engineering, Physics, or Chemistry, can be fruitfully partitioned into a number of disciplines, many of which present power law behavior. On the contrary, disciplines in the Biomedical Sciences and Clinical Medicine often fail to be represented by a power law. At any rate, higher aggregates for which the existence of a power law cannot be rejected tend to cover between 70 and 80% of all articles in the sample and, when they exist, power laws at these aggregate levels tend to be flatter, smaller and accountable for smaller percentages of citations than those at the sub-field level.

(7) It is important to emphasize that the considerable differences found in the power law characteristics at all aggregation levels go against the universality claim in Radicchi et al. (2008) at one key segment of citation distributions: the tip of the upper tail, or the place where citation excellence resides.

4 The evaluation of research units

4.1 The state of the art

It is illuminating to review how the specialists in bibliometrics address the evaluation of the scientific performance of research units. There are two types of output indicators. Firstly, there is the publication share during a given time period. Secondly, when there is information on the citations received by these publications, two other indicators are

typically added: the share of total citations, and some measure of the citation impact of the average paper. When the only information assumed to be available is the homogeneous field to which each unit's publications belong and the number of citations they receive, the MCR is a good overall indicator of scientific performance.[6]

Consider the important case of the comparison between the US and the EU. Three methodological points should be noted at the outset. Firstly, TS distinguishes 22 fields comprising 20 broad fields for the natural sciences and two for the social sciences. Although this firm does not provide a link between the 219 WoS categories and the 22 broad fields, TS assigns each article in our dataset to a single broad field. Given the illustrative nature of our work at this point (Albarrán et al. 2010, 2011b,c), in this Section we work at this high aggregate level, and assume that these 22 fields are homogeneous. In this way, the thorny problems discussed in Sect. 3.1 about the multiple assignments of articles to WoS categories, as well as the difficulties involved in the aggregation from the WoS to other levels, are provisionally avoided. Secondly, in every co-authored article by people working in a US and a European research center a whole count is credited to each contributing geographical area. Only domestic articles, or articles exclusively authored by one or more scientists affiliated to research centers either in the US or the EU alone, are counted once. Consequently, when the 1998–2002 dataset is partitioned into the US, the EU, and the RW the total number of articles in such extended count is 13.6% more than the standard count in which all articles are counted once. Thirdly, although the TS database covers 36 languages, there is a general agreement that it suffers from an English language bias. Some might argue that for the Social Sciences other than Economics, and perhaps also for Psychology and Psychiatry and the Behavioral Sciences, the TS database favors the US versus the EU. However, the remarkable findings by van Leeuwen et al. (2001) for the case of the life and other natural sciences indicate that countries such as Germany, France, and Switzerland have a decreasing though still significant number of publications in non-English journals that have a considerably lower impact than the English-language journals. Thus, when the publications in these non-English journals, but not their citations to articles in English, are removed from the publication output, the impact score of these countries shows increases above at least 10%. On the other hand, taking into account that English can be considered the international language of science, we follow the usual practice of using the TS data under the reasonable assumption that "*the international journal publications in these databases provide a satisfactory representation of internationally accepted* (*'mainstream'*) *research, especially high-quality 'laboratory based' basic research in the natural sciences, medical sciences, and life sciences conducted in the advanced industrialized nations*" (EC 2003, p. 439).

Since the mid 1990s, the EU publication share is greater than that of the US in a majority of scientific fields as well as in all fields combined. Analysts working for the European Commission, unduly impressed by this fact, have developed the "European Paradox" notion—popularized in the *First European Report on Science and Technology Indicators* (EC 1994)—according to which Europe plays a leading world role in terms of scientific excellence but lacks the entrepreneurial capacity of the US to

[6] See Albarrán et al. (2011a,c) for references to the case when there is information about the journals where the publications appear.

transform it into innovation, growth, and jobs (see Delanghe et al. 2011). Apparently, the problem lies not in the EU's scientific performance but elsewhere. Similarly, within the academic literature many papers transmit an optimistic view about the EU's performance. For instance, in his influential contribution, King (2004) states that "*the EU now matches the United States in the physical sciences, engineering and mathematics, although still lags in the life sciences*". But this statement refers to the share of total citations in these fields whose size depends on the corresponding publication shares that are generally greater in the EU. However, once the number of articles is also taken into account the MCR in all fields become greater in the US (for a literature review and some evidence about standard indicators, see Section II and Table 2, respectively, in Albarrán et al. 2010).

As indicated in the Introduction, a single statistic of centrality, such as the MCR or the median, may not adequately capture the skewness of distributions. There are several ways of taking into account this feature. Here we will refer to two alternatives. Firstly, among the battery of indicators used in one of the most influential research centers in bibliometrics, the Centre for Science and Technology Studies at Leiden University, one possibility is to complete the unit's MCR in a given field with the percentage of uncited papers, and the percentage contribution to the top 5 or 1% of highly cited papers (see, *inter alia* Moed et al. 1985; Moed and van Raan 1988; Moed et al. 1995; van Raan 2004; Tijssen et al. 2002, and van Leeuwen et al. 2003). Secondly, in Albarrán et al. (2010) we evaluate the performance of the US and the EU by comparing their publication shares at a large number of percentiles p of the world citation distribution. When $p = 0.1$, for example, the shares refer to the set of articles after discarding the 10% least cited, or to 90% of the most highly-cited articles. For a given geographical area, the graph of the publication shares as p increases from 0.1 to 0.2, 0.3, etc., reflects its relative performance as the publications impact measured by the number of citations increases. The comparisons of such graphs in Fig. 2 for the two geographical areas in a number of selected fields and all fields combined provide an eloquent picture of their relative situation at many points of the citation distribution.

Figure 2 deserves three comments. (i) The EU share of total publications in all sciences in 1998–2002 is about 4% greater than that of the US. However, as soon as we turn from the sheer production of scientific articles toward the impact they have in terms of total citations received during the entire period 1998–2007, the relative situation of the two geographical areas is dramatically reversed: for all sciences taken together, the US publication share becomes greater than the EU's for the top 50% of the most highly cited articles. (ii) Except for Agriculture Sciences, in the remaining 21 fields the dominance of the US over the EU among the most influential articles is overwhelming: the EU publication share is surpassed by the US share for all percentiles beyond the top 45% or the top 4% of the most highly cited articles, depending on the case. Interestingly enough, Economics and Business is the field where the US dominance is the greatest. (iii) The US curves tend to have a positive slope and, when the upper tail is reached at $p = 0.90$, they all clearly rise without exception. However, in about 10 fields the EU share remains relatively flat or slightly increases, while in the remaining 12 decreases at that crucial stage.

The overall conclusion is inescapable. Independently of sectoral details, there is a large gap between the international impact achieved by the US and the EU (Dosi et al.

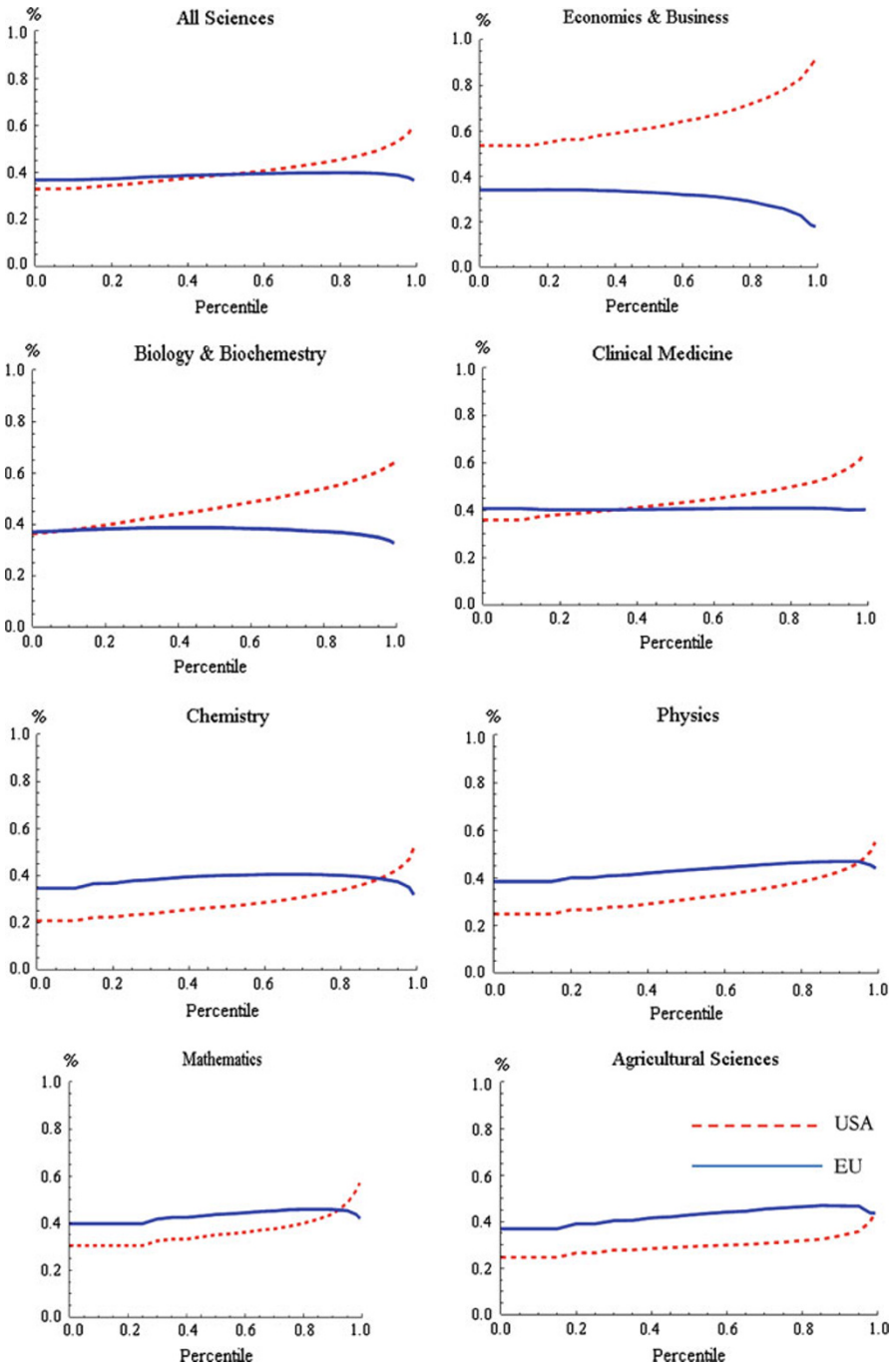

Fig. 2 Publication shares at different percentiles of citation distributions. 1998–2002 Articles with citations received during 1998–2007 in selected fields

2006; Leydesdorff and Wagner 2009, reach the same conclusion based on much more limited evidence).

4.2 A new proposal inspired by poverty measurement

As indicated in the Introduction, the skewness of citation distributions leads to their description in terms of two indicators: a high- and a low-impact indicator defined over the set of articles with citations above or below a reasonable CCL. Which indicators should be used in practice? Foster and Shorrocks (1991) show that the ranking of citation distributions induced by a family of low-impact indicators—the FGT family originally suggested by Foster et al. (1984)—is essentially characterized in terms of a number of interesting properties. These same properties lead to the selection of an equally convenient class of FGT high-impact measures that is the counterpart of the family just mentioned. Members of the FGT families capture different dimensions of the phenomena to be measured. The first member of each family coincides with the low- or the high-impact percentage of papers, measuring what was referred in the Introduction as the *incidence* aspect of the two phenomena under investigation. The second member incorporates as well as a measure of the aggregate gap between the actual number of citations received by each low- or high-impact paper and the CCL, that is, a measure of the *intensity* of the phenomena in question. Finally, in addition, the third member of the families is sensitive to the citation *inequality* in the sense that an increase in the coefficient of variation increases both the low- and the high-impact measures. Instead, the alternatives reviewed in Sect. 4.1 only capture some of these dimensions. The percentage of papers at one or several percentiles of the world citation distribution only measures the incidence aspect, while the MCR itself is silent about the distributive characteristics on either side of the mean (see Albarrán et al. 2011a, for a full discussion of the properties satisfied by our pair of indicators and the alternatives found in the bibliometrics literature).

Of course, whether the properties enjoyed by the FGT indicators are of any interest is not merely a formal issue. The value added, if any, can only be revealed by their use in practice. In what follows, we will briefly present some of the results in Albarrán et al. (2011b) that contain the first application of the new methodology to articles published by the US, the EU and the RW in 1998–2002, with a five year citation window, and with the CCL fixed at the 80th percentile of the world citation distribution in each of 22 TS scientific fields. Both families of FGT indicators are additively decomposable in the sense that, for any partition of a citation distribution, the overall high-impact level, for example, can be expressed as the weighted average of all sub-group high-impact levels, with weights equal to each sub-group publication shares. Then, the ratio of a sub-group index to the world index is greater than, equal to, or smaller than one whenever the sub-group contribution to the overall world level is greater than, equal to, or smaller than the sub-group publication share. The information about these ratios for every field, and the three members of the FGT families of high- and low-impact indicators is in Table 4 in Albarrán et al. (2011b). For the six fields already selected in Fig. 2, the results about the high-impact measurement are illustrated in Fig. 3. The US, the EU, and the RW appear in Fig. 3 from left to right. In each field and each area,

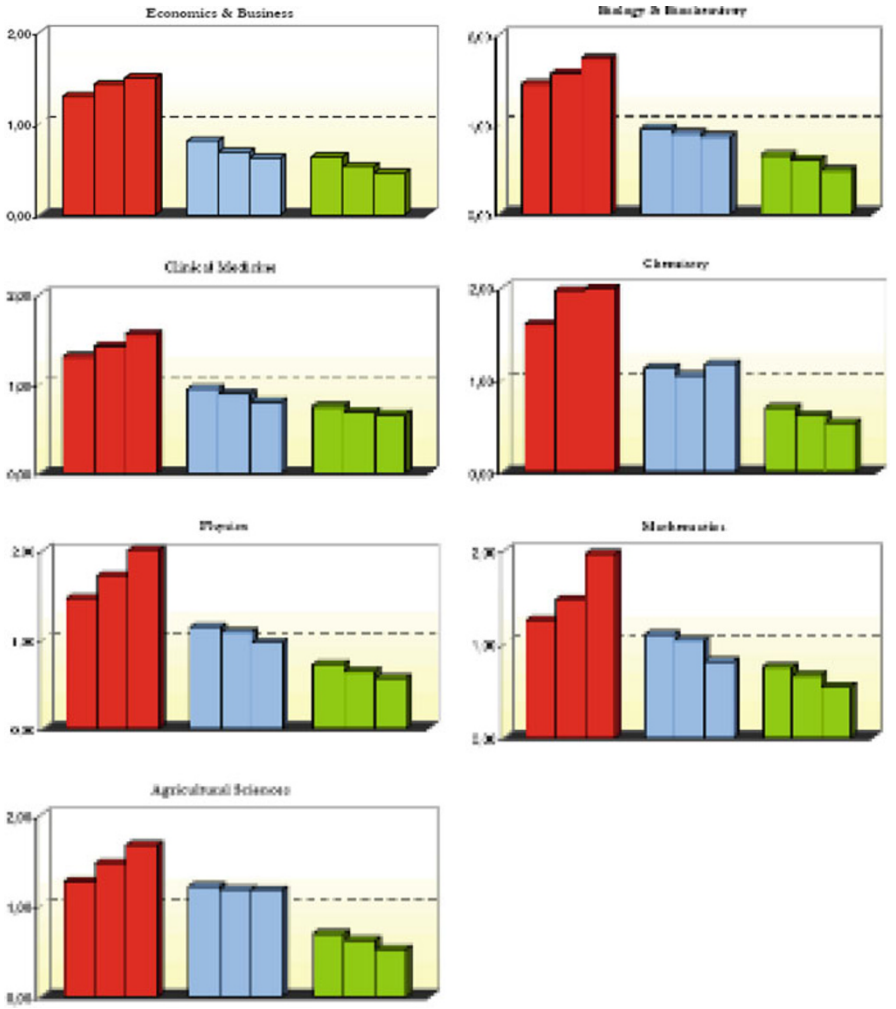

Fig. 3 The relative contribution to world high-impact levels by the US, the EU, and the RW (from *left to right*) according to incidence, intensity, and citation inequality members of the FGT of high impact indicators. 1998–2002 Articles with a five-year window in selected fields

the three bars in Fig. 3 reflect the ratios of an area index to the world index for each of the first three members of the FGT family of high-impact indicators.

(1) The US occupies a truly enviable position: as we keep introducing new measurement dimensions the ratio of the US to the world index strictly increases in all fields, indicating that its observed contribution to the world high-impact level is always greater than what is expected from its publication share. Essentially, the RW presents the opposite pattern: these ratios systematically decrease as new dimensions are taken into account. Finally, the EU high-impact performance is not very impressive. The EU ratios continuously decrease as we proceed from the incidence to the intensity and the citation inequality dimensions.

(2) In connection with the importance given in some quarters to publication shares, it should be emphasized that the absolute number of articles authored by the RW is considerably larger than that of the EU or the US in 13 fields. In turn, as we know, more articles are written in the EU than in the US in 14 fields. Given the results just summarized, this indicates that an area's large publication share within a field is no guarantee at all of a good high-impact performance. At the same time, publication efforts across fields within a geographical area are also unrelated to good high-impact performances.

(3) For reasons of space we will simply mention the following two points. Firstly, as far as the high-impact is concerned, raising the CCL from the 80th to the 95th percentile of the world citation distribution does not dramatically alter the relative situation of geographical areas and/or scientific fields. Secondly, as has been repeatedly observed in the bibliometrics literature, we find that international co-authorship as a whole is vastly successful (see Section 4.3, 4.4 in Albarrán et al. 2011b).

(4) Finally, it is important to know how this approach fares versus the alternatives. In Albarrán et al. (2011c), the results obtained with the new methodology are compared with those that can be obtained for each geographical area using the alternative with better properties, namely, what we call the Leiden triad of indicators, consisting of the MCR, and the area's percentage contribution to the set of uncited papers and to the top 5% of highly-cited papers in a given field. In brief, it is found that from an ordinal point of view following the Leiden or the new approach produces extremely similar results. However, considerable differences arise when the aim is the cardinal comparison of each area's relative situation. For reasons of space we will restrict ourselves to a single example: the differences between the results obtained with our preferred high-impact indicator and the MCR are greater than 20% half of the time. In particular, under the MCR criterion the US situation systematically worsens, while the relative situation of the EU and, above all, the RW appears reinforced. Thus, the question boils down to the following choice: to complete the MCR with percentage indicators of what happens at both tails of a citation distribution, or to use an integrated framework in which any citation distribution can be conveniently described by a pair of high- and low-impact indices whose properties not only have been extensively discussed in the axiomatic literature but appear to be useful in the empirical work, and admit a number of extensions which will be briefly mentioned in the concluding Sect. 5.

5 Conclusions and extensions

(1) Using a large dataset we have presented convincing systematic evidence about the existence of fundamental regularities in the shape of reference and citation distributions at different aggregation levels. As Lehmann et al. (2003) eloquently summarize: "*The picture which emerges is thus a small number of interesting and significant papers swimming in a sea of dead papers*" (p. 7). This is important because, regardless of the myriad of motives guiding specific citations by researchers, we are confronted with a social institution that calls for a single theoretical explanation of the decentralized process whereby scientists make references that a few years later will translate into a highly skewed citation distribution crowned in many cases by a power law.

Recent contributions using a social network approach by, for example, Dorogovstev and Mendes (2001); Jackson and Rogers (2007), and Peterson et al. (2010) constitute a formidable first attempt in this direction. The similarities that have been documented about citation distributions seem to indicate that a plausible working hypothesis is that the distribution of talent to achieve an international impact is, certainly skewed, but similar in all sciences.

Nevertheless, it should be noted that the study of whether a power law cannot be rejected is only a first step (see the discussion in Albarrán and Ruiz-Castillo 2011, and Albarrán et al. 2011d). New tests must be applied confronting power laws with alternative distributions, confidence intervals for the power law parameters must be estimated, robust estimation methods to the presence of extreme observations must be explored, and appropriate statistical models for the entire citation distribution must be tried out.

(2) It has been observed that, in spite of broad similarities among citation distributions, mean normalization at the sub-field level does not lead to a universal distribution. Two points should be noted in this respect. Firstly, differences at the lower tail, including the percentage of articles without citations, may partially depend on the fact that we have taken a common citation window for all sub-fields. Citation windows should be set so that the citation process that works at different speeds across sub-fields reaches the same stage in all of them. Variable citation windows may strengthen the similarities at the lower tail. Secondly, it has been emphasized that, at the tip of the upper tail, at least some citation distributions exhibit considerable differences. This seems to preclude the comparability of the citation impact of articles in different sub-fields. However, the existing regularities at other segments of the distribution might be enough for practical purposes. Let $p \in (0, 100)$ be a certain percentile of any citation distribution, and let $c_i(p)$ be the corresponding number of citations in field i, $i = 1, \ldots, I$. Let us define *adjustment factors* for all fields in terms of one of them, say field I, as follows: $f_i(p) = c_i(p)/c_I(p)$ for all $i \neq I$. We have already seen that means are generally reached at about the 70th percentile. Therefore, one way to assess whether citation distributions are at a comparable distance is to compare adjustment factors at different percentiles in the range $p \in [70, 100)$. Excluding the Multidisciplinary field for its intrinsic peculiarities, preliminary results for the remaining TS indicate that this is indeed the case for $p = 70$, 80, 85, 90, and 95. In the metaphor according to which citation distributions can be identified with income distributions, adjustment factors are seen as *exchange rates* that serve to express the citations received by articles in different fields in the same currency with a tolerable margin of error.

(3) As reviewed in Sect. 4, there is a strong case for using two statistics to summarize the typical shape of citation distributions, one low- impact index akin to an economic poverty index, and one high-impact index akin to some sort of affluence indicator. This approach can be extended in a number of ways. Firstly, this framework can be profitably used for the analysis of inter-temporal trends. Recall that, for any partition, overall high- or low-impact levels can be expressed as the weighted average of each subgroup's high- or low-impact levels, where the weights are the subgroups' publication shares. Therefore, inter-temporal comparisons of overall levels can be accounted for by changes in publication shares and by changes in subgroups' index values. Consider the case of an emergent country like China, whose scientific performance has

been recently quickly improving. This approach would allow distinguishing between the relative importance of increasing publication shares or of improvements in performance according to high- or low-impact indicators. Secondly, the first empirical application of this methodology has been based on a choice of two convenient CCLs, and a number of indicators with useful properties for applied work. However, results on economic poverty dominance should help us search for high- or low-impact comparisons robust to the choice of the CCL and the selection of indicators in a wide class of admissible ones (see *inter alia* Jenkins and Lambert 1997).

(4) The first empirical application of this approach has studied a simple partition of the world into three large geographical areas for the 22 TS broad fields. Three interesting extensions seem possible. Firstly, it is important to replicate the analysis at the level of the 219 WoS categories in our dataset. Provisional results indicate that the EU has more publications than the US in 113 sub-fields. However, judging from the high-impact perspective, the EU is ahead of the US only in 30 out of 219 sub-fields. In 57 and 14 sub-fields within the 186 natural sciences and the 33 social sciences, respectively, the US has a high-impact indicator at least twice as large as the EU. Judging from the low-impact perspective, the EU situation is somewhat more favorable. For example, the EU is ahead—namely, it has a smaller low-impact level—in 56 out of 219 sub-fields. Nevertheless, for all sciences as a whole the US low-impact indicator is 12.3% smaller than that of the EU.

Secondly, in line with the evaluation tradition mentioned in the Introduction, any move from the sub-field to higher aggregate levels should take into account scale differences across heterogeneous sub-fields. Focusing only on the high-impact gap, provisional results indicate that normalization does not systematically favor any of the two geographical areas (for example, normalization favors the US in 49 out of 80 disciplines and the EU in 29). After normalization, the EU is ahead or at the same level in only five disciplines, while the US dominates the EU by more than 100% in 33. On the other hand, although normalization reduces the US/EU high-impact gap by a non-negligible 16.8% in all sciences as a whole, the US high-impact indicator at this level is about 61% greater than that of the EU.

Thirdly, it can be concluded that, judging from citation impact, the so-called European Paradox hides a truly *European Drama*: the dominance of the US over the EU in the basic and applied research published in the periodical literature, before and after normalization, is overwhelming at all aggregation levels. The analysis, of course, might be extended in rather obvious directions towards specific countries within the EU and the RW, and even towards individual research centers.

(5) As indicated in the Introduction, it would be interesting to evaluate research units in terms of the h-index, an indicator of excellence with very different properties from our high-impact index. To begin with, one needs to build a homogeneous field out of a set of heterogeneous sub-fields. Scale normalization along the lines already discussed should allow us to compare the h-index of two research units of the same size. The remaining difficulty is the incomparability of the normalized h-index of two research units of different size. Following Molinari and Molinari (2008a,b), ongoing research suggests the following bootstrap procedure. Consider the normalized field distribution, as well as the number of articles published by a certain research unit in that field. Select a large number of random samples of that size in the normalized distribution.

The empirical distribution of the h-index in those samples should provide an adequate benchmark for the actual normalized h-index of the research unit in question. For example, consider the percentile of the distribution in which the actual normalized h-index is placed. Doing the same for a research unit of a different size, the two percentiles should be normatively comparable.

Note that, in principle, this approach might be also applied to the all-sciences case. This might make possible the following two important exercises. Firstly, the comparison of research units—such as Universities or countries—working simultaneously in several, or all scientific fields. Secondly, consider the allocation of a certain monetary sum among a set of scientific fields. The proportional and the egalitarian allocations constitute two sensible solutions to this problem. However, it is interesting to search for an alternative between these two on the grounds that scientific excellence can be expected to vary positively with field size, but in a less than proportional way—a property satisfied by the h-index. For each field size, consider the possibility of estimating the mean of the distribution of normalized h-indices in the random samples of that size. A possible solution is the allocation of the money to the fields in proportion to the mean normalized h-index for each field size.

(6) Finally, the idea of the identification of a citation distribution with an income distribution can be reinterpreted in a longitudinal context: a given crop of articles published on a certain date that receives citations in yearly waves gives rise to a panel data sample. Then, three issues can be explored. Firstly, the dynamic features unveiled by a statistical model that explains the probability that an article will be cited should be reckoned with by any positive theory of the citation process of the type mentioned in the first paragraph of this Section. Secondly, so far all the normalization procedures we have discussed have been unconditional procedures that aim to eliminate the scale differences that separate citation distributions, no matter what the forces that may have caused them are. The estimation of a dynamic model may pave the way to (conditional) normalization procedures that, among the variables accounting for scale differences, only consider those that are deemed to be relevant from a normative point of view. Finally, we can simply ask for the amount of time necessary for citation distributions to acquire their typical shape—an empirical question of obvious interest for those in charge of the evaluation of scientific research through citation analysis. As new citation waves arrive, re-rankings of articles in terms of citations received will take place. At the same time, any wave's citation inequality may be affecting the aggregate citation distribution formed by all citation waves produced so far. In this context, income mobility indices may help quantify the amount of "citation mobility" involved at each step in this process. In particular, we may learn when citation distributions acquire their typical shape by observing when citation mobility indices become stabilized.

To conclude, welfare economics and statistical analysis have been shown to have a relevant role in citation analysis, a promising research field made possible by the current availability of data that has unveiled the characteristics of citation patterns in vastly different scientific fields.

References

Albarrán P, Ruiz-Castillo J (2011) References made and citations received by scientific articles. J Am Soc Inf Sci Technol 62:40–49

Albarrán P, Crespo J, Ortuño I, Ruiz-Castillo J (2010) A comparison of the scientific performance of the US and Europe at the turn of the 21st century. Scientometrics 85:329–344

Albarrán P, Ortuño I, Ruiz-Castillo J (2011) The measurement of low- and high-impact in citation distributions: technical results. J Informetr 5:48–63

Albarrán P, Ortuño I, Ruiz-Castillo J (2011) High- and low-impact citation measures: empirical applications. J Informetr 5:122–145

Albarrán P, Ortuño I, Ruiz-Castillo J (2011c) Average-based versus high- and low-impact indicators for the evaluation of citation distributions with. In: Research evaluation (forthcoming)

Albarrán P, Crespo J, Ortuño I, Ruiz-Castillo J (2011d) The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics 88:385–397

Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F (2009) h-Index: a review focused in its variants, computation and standardization for different scientific fields. J Informetr 3:273–289

Bornmann L, Daniel H-D (2008) What do citation counts measure? J Doc 64:45–80

Clauset A, Shalizi CR, Newman MEJ (2007) Power-law distributions In empirical data. SIAM Rev 51:661–703

Cole JR (2000) A short history of the use of citations as a measure of the impact of scientific and scholarly work. In: Cronin B, Atkins HBThe web of knowledge: a festschrisft in honor of Eugene Garfield. Information Today, Medford

Delanghe H, Sloan B, Muldur U (2011) European research policy and bibliometric indicators, 1990–2005. Scientometrics 87:389–398

Dorogovstev S, Mendes J (2001) Scaling properties of scale-free evolving networks: continuous approach. Phys Rev E 85:4633–4636

Dosi G, Llerena P, Sylos Labini M (2006) Science-Technology-Industry links and the 'European Paradox': some notes on the dynamics of scientific and technological research in Europe. Res Policy 35:1450–1464

EC (1994) First European Report on Science and Technology Indicators, Directorate-General XII, Science, Research, and Development. Luxembourg: Office for Official Publications of the European Community

EC (2003) Third European Report on Science and Technology Indicators, Directorate-General for Research. Luxembourg: Office for Official Publications of the European Community, http://www.cordis.lu/rtd2002/indicators/home.html

Egghe L (2005) Power laws in the information production process: Lotkaian informetrics. Elsevier, Amsterdam

Foster JE, Greeer J, Thorbecke E (1984) A class of decomposable poverty measures. Econometrica 52:761–766

Foster JE, Shorrocks A (1991) Subgroup consistent poverty indices. Econometrica 59:687–709

Glänzel W (2010) The application of characteristics scores and scales to the evaluation and ranking of scientific journals. In: Proceedings of INFO 2010, Havana, Cuba, pp 1–13 (forthcoming)

Glänzel W, Schubert A (2003) A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. Scientometrics 56:357–367

Hirsch J (2005) An index to quantify an individual's scientific research output. Proc natl Acad Sci USA 102:16569–16572

Jackson M, Rogers B (2007) Meeting strangers and friends of friends: how random are social networks? Am Econ Rev 97:890–915

Jenkins S, Lambert P (1997) Three 'I's of poverty curves, with an analysis of UK poverty trends. Oxf Econ Pap 49:317–327

King D (2004) The scientific impact of nations. Nature 430:311–316

Lehmann S, Lautrup B, Jackson AD (2003) Citation networks in high energy physics. Phys Rev E68:026113–026118

Leydesdorff L, Wagner C (2009) Is the US losing ground in science? a global perspective on the world science system. Scientometrics 78:23–36

Marchant T (2009) An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. Scientometrics 80:325–342

Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet Math 1:226–251
Moed HF, Burger WJ, Frankfort JG, van Raan AFJ (1985) The use of bibliometric data for the measurement of university research performance. Res Policy 14:131–149
Moed HF, van Raan AFJ (1988) Indicators of research performance. In: van Raan AFJ (ed) Handbook of quantitative studies of science and technology. North Holland, Amsterdam, pp 177–192
Moed HF, De Bruin RE, van Leeuwen ThN (1995) New bibliometrics tools for the assessment of national research performance: database description, overview of indicators, and first applications. Scientometrics 33:381–422
Molinari JF, Molinari A (2008) A new methodology for ranking scientific institutions. Scientometrics 75:163–174
Molinari A, Molinari J-F (2008) Mathematical aspects of a new criterion for ranking scientific institutions based on the h-index. Scientometrics 75:339–356
Newman MEJ (2005) Power laws, Pareto distributions, and Zipf's law. Contemp Phys 46:323–351
Peterson G, Presse S, Dill K (2010) Nonuniversal power law scaling in the probability distribution of scientific citations. PNAS 107:16023–16027
Quesada A (2009) Monotonicity and the Hirsch index. J Informetr 3:158–160
Quesada A (2010) More axiomatics for the Hirsch index. Scientometrics 82:413–418
Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: toward an objective measure of scientific impact. PNAS 105:17268–17272
Ruiz-Castillo J (2011) The evaluation of citation distributions. Working Paper 11–12, Economics Department, Universidad Carlos III
Schubert A, Glänzel W, Braun T (1987) Subject field characteristic citation scores and scales for assessing research performance. Scientometrics 12:267–292
Sen A (1976) Poverty: an ordinal approach to measurement. Econometrica 44:219–230
Tijssen R, Visser M, van Leeuwen T (2002) Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference. Scientometrics 54:381–397
Tijssen RJW, van Leeuwen TN (2003) Bibliometric analyses of world science, extended technical annex to Chapter 5 of the third European Report on science and technology indicators. mimeo, Leiden University
van Leeuwen T, Moed H, Tijssen R, Visser M, Raan Avan (2001) Language biases in the coverage of the science citation index and its consequences for international comparisons of national research performance. Scientometrics 51:335–346
van Leeuwen T, Visser M, Moed H, Nederhof T, Raan Avan (2003) The holy grail of science policy: exploring and combining bibliometric tools in search of scientific excellence. Scientometrics 57:257–280
van Raan AFJ (2004) Measuring Science. In: Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems, Kluwer, Dordrecht
van Raan AFJ (2005) Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. Scientometrics 62:133–143
Waltman L, van Eck NJ, van Raan AFJ (2011) Universality of citation distributions revisited. Center for science and technological studies, Leiden University, The Netherlands, mimeo, http://arxiv.org/pdf/1105.2934v1
Weingart P (2005) Impact of bibliometrics upon the science system: inadvertent consequences? Scientometrics 62:117–131
Woeginger G (2008) An axiomatic characterization of the Hirsch-index. Math Soc Sci 56:224–232
Woeginger G (2008) A symmetry axiom for scientific impact indices. J Informetr 2:298–303